

GOVT 6019: Midterm Practice Questions — Solutions

Q1

A research assistant sends you a memo with five claims. Each one contains a mistake. For each claim, (i) name the problem and (ii) explain in one or two sentences what is wrong or what the analyst should do instead.

- a) “We measure protest activity using reports from one national newspaper, which rarely covers rural areas. We recorded 3,000 events, so with this much data any undercount of rural protests will average out.”

Answer. **Systematic error (bias) does not shrink with sample size.** The newspaper misses rural protests in a consistent direction, so every additional event carries the same undercount. More data averages out *random* error, not systematic error. Fix: add sources that cover rural areas, or restrict the claim to what the newspaper covers.

- b) “States with more union members have higher minimum wages. This shows that union members support higher minimum wages.”

Answer. **Ecological inference: the data are about states, the claim is about individuals.** Higher minimum wages in union-dense states could reflect state-level politics (e.g. Democratic legislatures) that affect everyone, not union members’ own preferences. Individual-level data (e.g. a survey) are needed for the individual claim.

- c) “To get national turnout, we averaged the 50 states’ turnout rates and got 61.3 percent.”

Answer. **Average of rates vs. pooled rate.** A simple average gives Wyoming the same weight as California. National turnout is total votes divided by total eligible voters, which weights each by its size (each voter counts once).

- d) “Among incumbents who lost re-election, 80 percent had been involved in a scandal. So an incumbent involved in a scandal has an 80 percent chance of losing.”

Answer. **Confusing $P(\text{Scandal} \mid \text{Lost})$ with $P(\text{Lost} \mid \text{Scandal})$.** The 80 percent is the share of losers with scandals. The chance of losing given a scandal also depends on base rates: how many incumbents lose overall and how many have scandals. By Bayes’ rule, $P(L \mid S) = P(S \mid L)P(L)/P(S)$.

- e) “Our 95 percent confidence interval for the share of voters who support the measure is [.44, .50]. So there is a 95 percent probability that true support lies between 44 and 50 percent.”

Answer. **Misinterpreting a confidence interval.** The true share is fixed; this interval either contains it or does not. The 95 percent describes the procedure: across many samples, about 95 percent of intervals built this way would contain the true share.

Q2

Every bill in a state legislature is referred to exactly one of three committees. Half of all bills go to Finance, 30 percent to Health and 20 percent to Education. Finance reports out (passes) 40 percent of its bills, Health 60 percent and Education 50 percent.

- a) What share of all bills pass committee? Show your work.

Answer. Law of total probability: $P(\text{Pass}) = .5(.4) + .3(.6) + .2(.5) = .20 + .18 + .10 = .48$.

- b) Compute $P(\text{Finance} \mid \text{Pass})$ and $P(\text{Pass} \mid \text{Finance})$. In words, what does each measure?

Answer. $P(\mathbf{Pass} \mid \mathbf{Finance})$. This is given directly in the question: Finance passes 40 percent of its bills, so $P(\text{Pass} \mid \text{Finance}) = .4$. It is the share of Finance bills that pass.

$P(\mathbf{Finance} \mid \mathbf{Pass})$. Use Bayes' rule, step by step.

1. Write the definition: $P(\text{Finance} \mid \text{Pass}) = \frac{P(\text{Finance} \cap \text{Pass})}{P(\text{Pass})}$.
2. Numerator, by the multiplication rule: $P(\text{Finance} \cap \text{Pass}) = P(\text{Pass} \mid \text{Finance}) P(\text{Finance}) = .4 \times .5 = .20$.
3. Denominator, from part (a): $P(\text{Pass}) = .48$.
4. Divide: $\frac{.20}{.48} = \frac{20}{48} = \frac{5}{12} \approx .42$.

It is the share of passed bills that came from Finance.

Check with counts. Out of 100 bills, 50 go to Finance and $.4 \times 50 = 20$ of them pass. In total, $20 + 18 + 10 = 48$ bills pass. So 20 of the 48 passed bills came from Finance: $20/48 = 5/12$.

- c) Are committee and passing independent? Use one numerical comparison.

Answer. No. $P(\text{Pass} \mid \text{Health}) = .6 \neq .48 = P(\text{Pass})$. (Any committee works: Finance .4 or Education .5 vs. .48.)

- d) A reporter writes: "Finance produces more passed bills than any other committee, so it is the easiest committee to get a bill through." Evaluate this claim using your answers above.

Answer. Wrong. Finance supplies the largest share of passed bills ($5/12$) only because it receives half of all bills. Its pass *rate*, .4, is the lowest of the three. The reporter confused $P(\text{Finance} \mid \text{Pass})$ with $P(\text{Pass} \mid \text{Finance})$; ease of passage is the second.

Q3

A candidate holds two town halls in different cities. At each, protesters show up with probability .3, independently of the other. Let X be the number of town halls with protesters.

- a) Write out the PMF of X .

Answer. $P(X = 0) = .7 \times .7 = .49$; $P(X = 1) = .3(.7) + .7(.3) = .42$; $P(X = 2) = .3 \times .3 = .09$.

- b) Check that your PMF is valid, and compute $P(X \leq 1)$.

Answer. All values in $[0, 1]$; $.49 + .42 + .09 = 1$. $P(X \leq 1) = F(1) = .49 + .42 = .91$.

- c) Compute $V[X]$. (Hint: $V[X] = E[X^2] - (E[X])^2$.)

Answer. $E[X] = 0(.49) + 1(.42) + 2(.09) = .6$. $E[X^2] = 0(.49) + 1(.42) + 4(.09) = .78$. $V[X] = .78 - .6^2 = .78 - .36 = .42$.

- d) Security costs \$5,000 for each town hall with protesters, plus a fixed \$2,000 for the campaign's security plan. Find the variance and the standard deviation of the security cost. You may leave square roots unevaluated.

Answer. $C = 5,000X + 2,000$. Adding a constant does not change spread, and multiplying by a multiplies the variance by a^2 : $V[C] = 5,000^2 \times .42 = 25,000,000 \times .42 = 10,500,000$. $\text{SD}[C] = 5,000\sqrt{.42}$, about \$3,240.

- e) Suppose instead both town halls are in the same city on the same weekend. Is independence still plausible? How would $V[X]$ change, and why?

Answer. Probably not: the same activists can attend both, so protest at one makes protest at the other more likely (positive covariance). Writing $X = X_1 + X_2$, $V[X] = V[X_1] + V[X_2] + 2 \text{Cov}(X_1, X_2)$, so a positive covariance makes $V[X]$ *larger* than .42: outcomes bunch at 0 and 2.

Q4

In County X, 2 percent of in-person provisional ballots and 10 percent of mail ballots were rejected. The county's overall rejection rate across both types was 4 percent.

- a) What share of County X's ballots were mail ballots? Show your work.

Answer. The overall rate is a weighted average. With w the mail share: $.02(1 - w) + .10w = .04 \Rightarrow .02 + .08w = .04 \Rightarrow w = .25$.

- b) County Y rejected 1 percent of in-person ballots and 8 percent of mail ballots. Half of its ballots were mail. Compute County Y's overall rejection rate.

Answer. $.5(.01) + .5(.08) = .005 + .04 = .045$.

- c) County Y has a lower rejection rate than County X for *each* ballot type, yet a higher overall rate. Explain how.

Answer. Mail ballots are rejected much more often than in-person ballots in both counties. Half of Y's ballots are mail, against a quarter of X's, so Y's overall rate puts more weight on the high-rejection type. The reversal comes from the different mix of ballot types (Simpson's paradox), not from Y doing worse within either type.

- d) County X counted 9,000 ballots and County Y 1,000. A reporter averages the two overall rates and reports a combined rate of 4.25 percent. What share of all 10,000 ballots was rejected? Why does it differ from the reporter's number?

Answer. Rejected: $.04(9,000) + .045(1,000) = 360 + 45 = 405$, so the pooled rate is $405/10,000 = .0405$. The reporter's average, $(.04 + .045)/2 = .0425$, gives the small county the same weight as the large one; the pooled rate counts each ballot once.

- e) A third county, Z, reported 0 ballots returned and -99 rejected, where the codebook says -99 means "data not available." What does this code print for County Z, and what should the analyst have done first?

```
d <- tibble(county = c("X", "Y", "Z"),
             returned = c(9000, 1000, 0),
             rejected = c(360, 45, -99))
d |> mutate(rate = rejected / returned) |> pull(rate)
```

Answer. It prints 0.040 0.045 $-\text{Inf}$: $-99/0$ is a negative number divided by zero, which R returns as $-\text{Inf}$, and any mean of this column would be $-\text{Inf}$. The analyst should first recode -99 to NA (e.g. `if_else(x < 0, NA, x)`) and drop counties with 0 ballots returned, since a rate is undefined there.

Q5

A simple random sample of 100 adults in a city records how many minutes of local news each watched yesterday. The sample mean is 22 minutes. The researcher reports a 95 percent confidence interval of 22 ± 3.9 , or about $[18.1, 25.9]$, using the multiplier 1.96.

```
> 3.9 / 1.96
[1] 1.989796
> 2 * pnorm(-1)
[1] 0.3173
```

- a) What is the estimate of the population mean, and what is the estimand?

Answer. Estimate: 22 minutes. Estimand: the population mean minutes of local news, μ , for all adults in the city. (The estimator is the rule, the sample mean $\bar{X}$.)

- b) Find the standard error. Round to a whole number.

Answer. The half-width is $1.96 \times \text{SE} = 3.9$, so $\text{SE} = 3.9/1.96 \approx 2$.

- c) Test $H_0 : \mu = 20$ against a two-sided alternative at $\alpha = .05$. Give the test statistic and your conclusion.

Answer. $t = (22 - 20)/2 = 1$. Two-sided $p = 2\Phi(-1) = .317 > .05$: fail to reject. Consistent with the interval, which contains 20.

- d) Interpret the p -value from (c).

Answer. If μ were 20, about 32 percent of samples would give a mean at least as far from 20 as 22 is, in either direction. It is not the probability that H_0 is true.

- e) About how many adults would the researcher need to cut the interval's width in half? Explain.

Answer. About 400. $\text{SE} = \sigma/\sqrt{n}$, so halving the SE needs $\sqrt{n}$ to double, i.e. n to quadruple: $4 \times 100 = 400$.

- f) A city council member says: "So 95 percent of residents watch between 18 and 26 minutes of local news a day." What is wrong? Use the standard error to estimate the standard deviation of viewing time across individuals.

Answer. The interval is about the *mean*, not individual residents. Individual viewing varies far more: $s = \text{SE} \times \sqrt{n} \approx 2 \times 10 = 20$ minutes, so many residents watch 0 and others watch an hour.

- g) Viewing time is strongly right-skewed: most people watch little and a few watch a lot. Does this make the confidence interval invalid? Explain.

Answer. No. With $n = 100$, the CLT says the sampling distribution of $\bar{X}$ is approximately normal even though the individual data stay skewed.

- h) If 200 researchers each drew their own sample of 100 and built a 95 percent interval this way, about how many intervals would fail to contain μ ?

Answer. About $.05 \times 200 = 10$.

To estimate how much viewing time varies across residents, the researcher considers two estimators of the population variance σ^2 :

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2.$$

You may use the facts that $E[S^2] = \sigma^2$ and $E[\hat{\sigma}^2] = \frac{n-1}{n} \sigma^2$.

- i) Which estimator is unbiased? Is the other one too high or too low on average, and by what fraction when $n = 10$ and when $n = 100$?

Answer. S^2 is unbiased. $\hat{\sigma}^2$ is biased *downward*: on average it equals $\frac{n-1}{n} \sigma^2$, so it is too low by $\frac{1}{n}$ of σ^2 : 10 percent at $n = 10$, 1 percent at $n = 100$.

- j) Is $\hat{\sigma}^2$ consistent? Explain what happens to its bias as n grows.

Answer. Yes. Its bias, $-\sigma^2/n$, shrinks to zero as n grows because $\frac{n-1}{n} \rightarrow 1$, and its spread also shrinks, so it converges to σ^2 . Biased in small samples but consistent.

- k) In one or two sentences, give the intuition for why dividing by n underestimates σ^2 .

Answer. The deviations are measured from $\bar{X}$, which is computed from the same data and so sits closer to the observations than the true mean μ does. That makes $\sum (X_i - \bar{X})^2$ smaller on average than $\sum (X_i - \mu)^2$; dividing by $n - 1$ instead of n corrects for this.